


Challenges of Index Exchange

Djoerd Hiemstra
Gijs Hendriksen
Chris Kamphuis
Arjen P. de Vries

A SEARCH ENGINE COMPANY... IN THE YEAR 2023

 A horizontal search bar with a thin blue border. The right end of the bar is a solid blue rectangle containing a white magnifying glass icon.

A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web

 A search bar consisting of a white rectangular input field with a thin blue border, followed by a blue square button containing a white magnifying glass icon.


A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages

A search bar consisting of a white rectangular input field with a thin blue border, followed by a blue square button containing a white magnifying glass icon.

A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)

 A search bar consisting of a white rectangular input field with a thin blue border, followed by a blue square button containing a white magnifying glass icon.

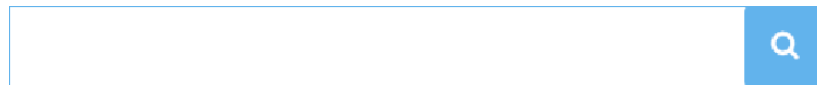
A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)

A horizontal search bar with a light blue border and a blue search button on the right side containing a magnifying glass icon.

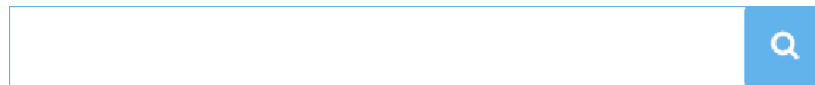
A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)
- Free web analytics (usage statistics)



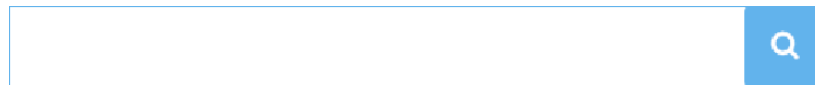
A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)
- Free web analytics (usage statistics)
- Sells advertisements



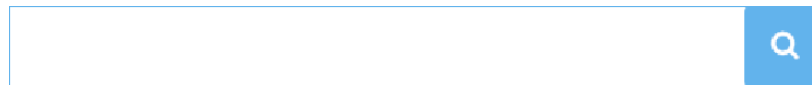
A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)
- Free web analytics (usage statistics)
- Sells advertisements
- Owns web browser



A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)
- Free web analytics (usage statistics)
- Sells advertisements
- Owns web browser
- Owns operating system



A SEARCH ENGINE COMPANY... IN THE YEAR 2023

- Crawls the web
- Indexes crawled pages
- Operates search engine (backend)
- Provides search application (frontend)
- Free web analytics (usage statistics)
- Sells advertisements
- Owns web browser
- Owns operating system



WHAT WE NEED



WHAT WE NEED

1) Regulation



WHAT WE NEED

1) Regulation



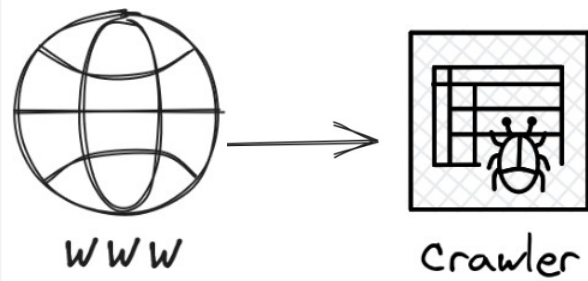
2) Open standards



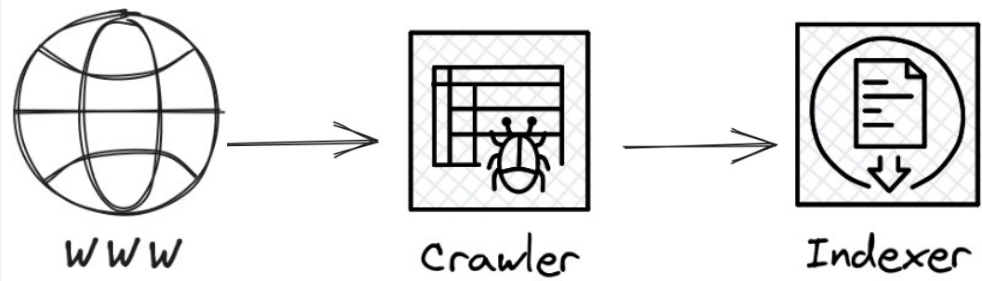
BUILDING SEARCH ENGINES COLLABORATIVELY



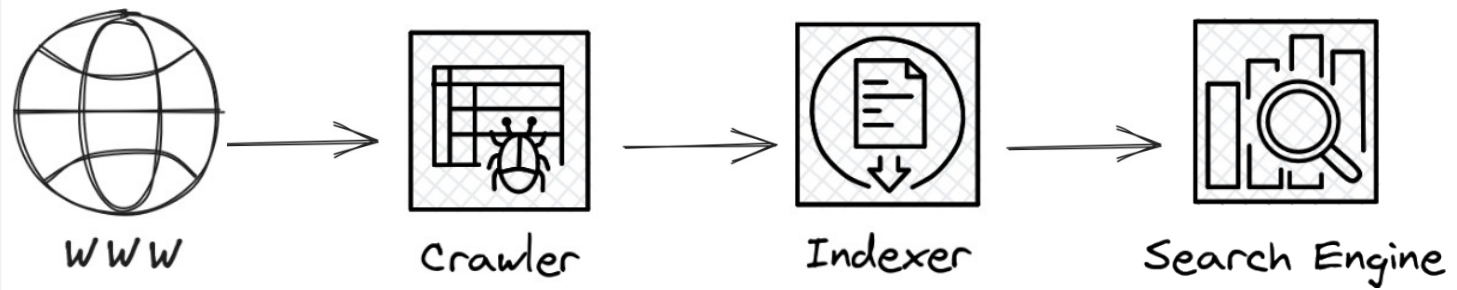
BUILDING SEARCH ENGINES COLLABORATIVELY



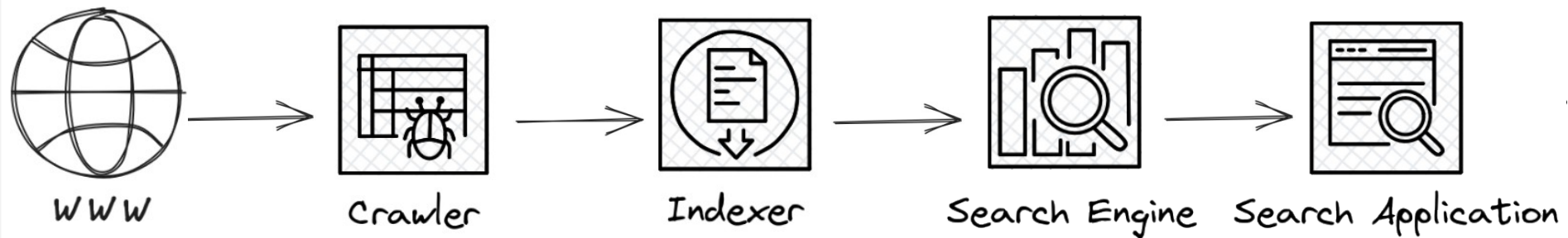
BUILDING SEARCH ENGINES COLLABORATIVELY



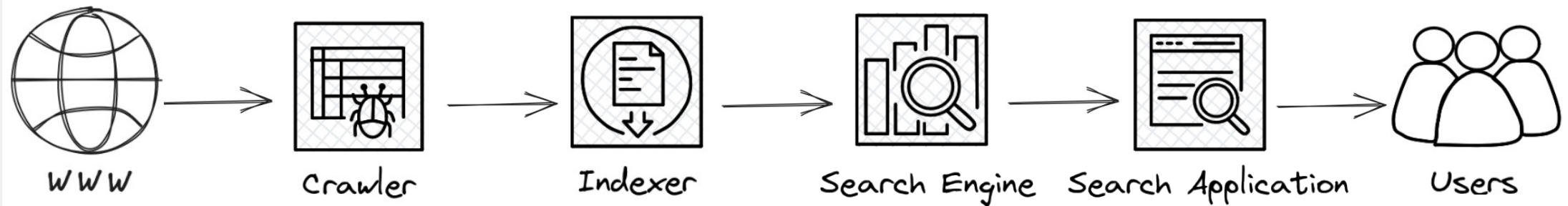
BUILDING SEARCH ENGINES COLLABORATIVELY



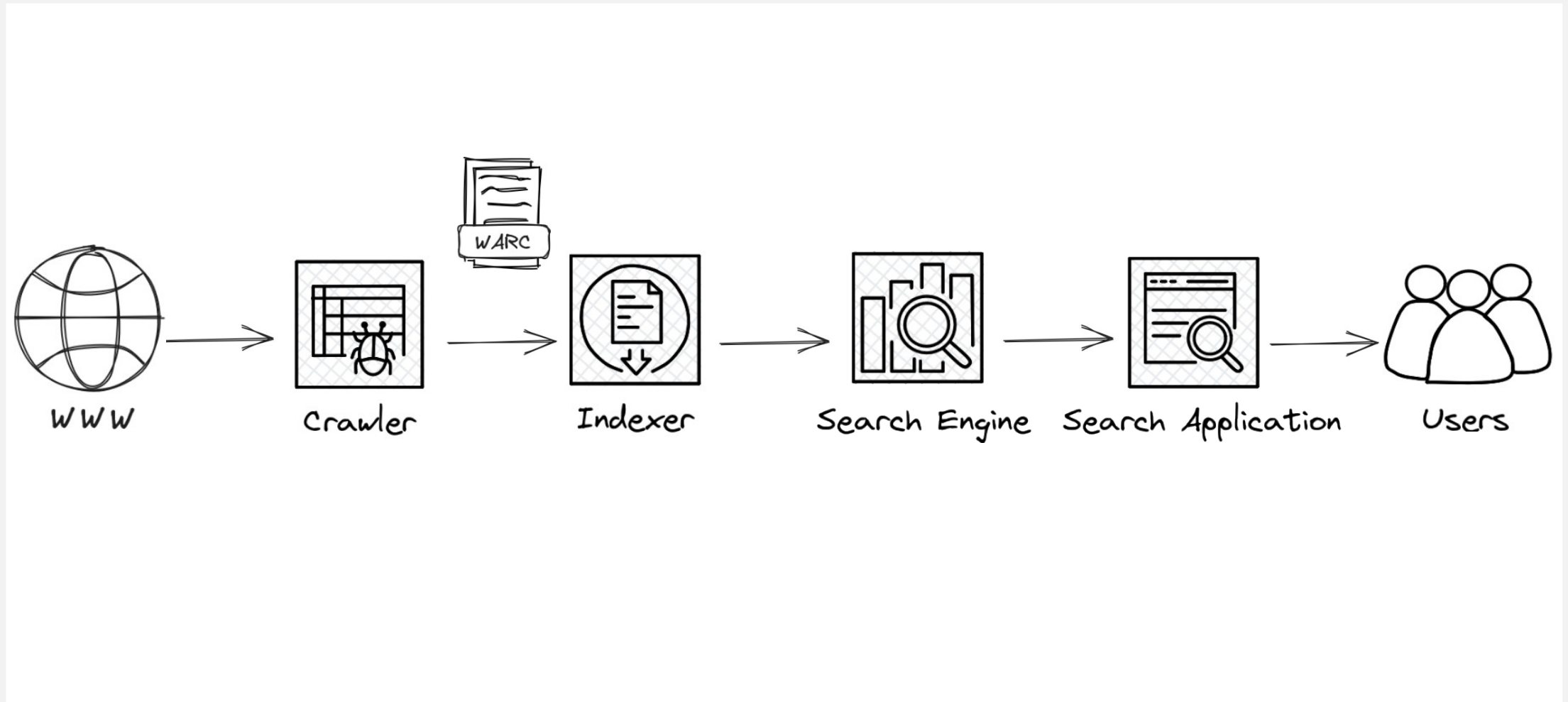
BUILDING SEARCH ENGINES COLLABORATIVELY



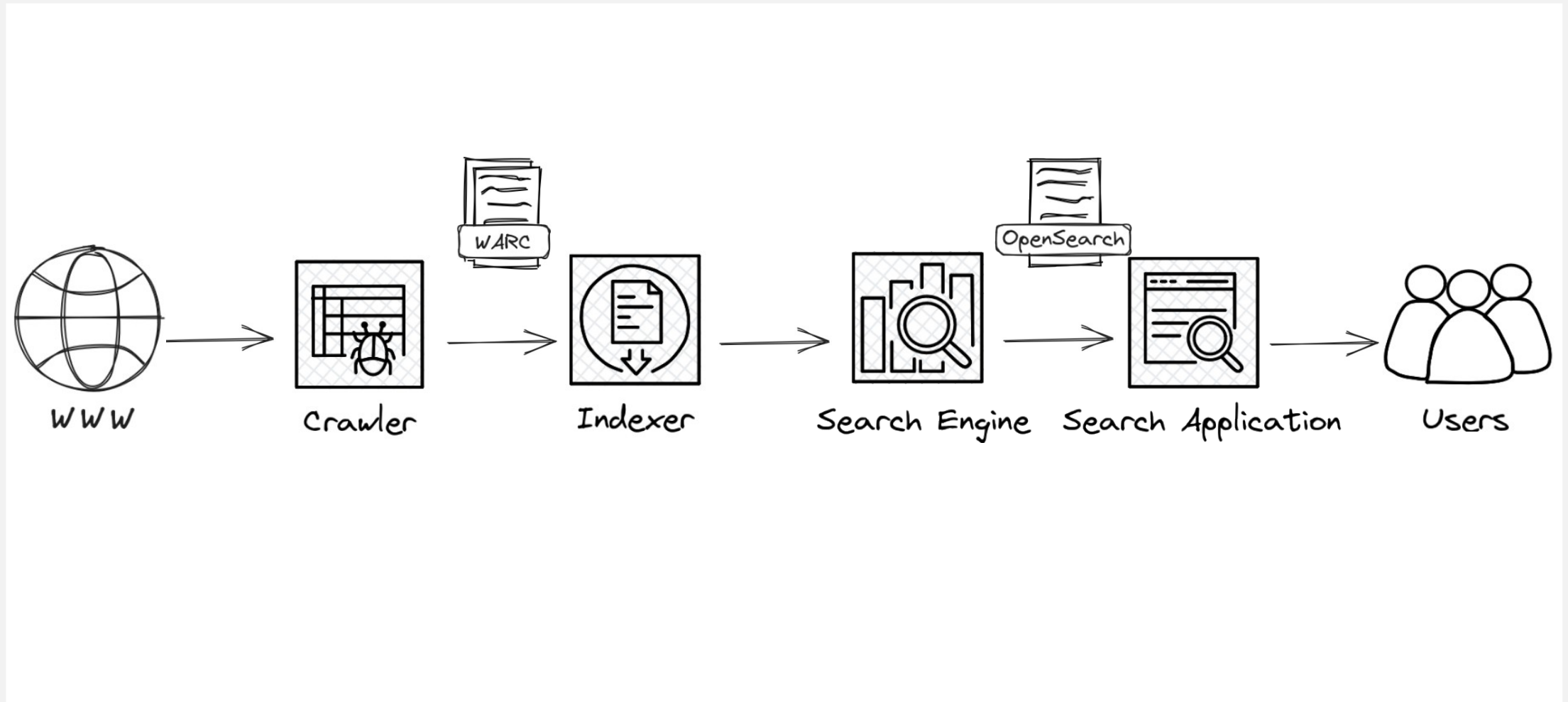
BUILDING SEARCH ENGINES COLLABORATIVELY



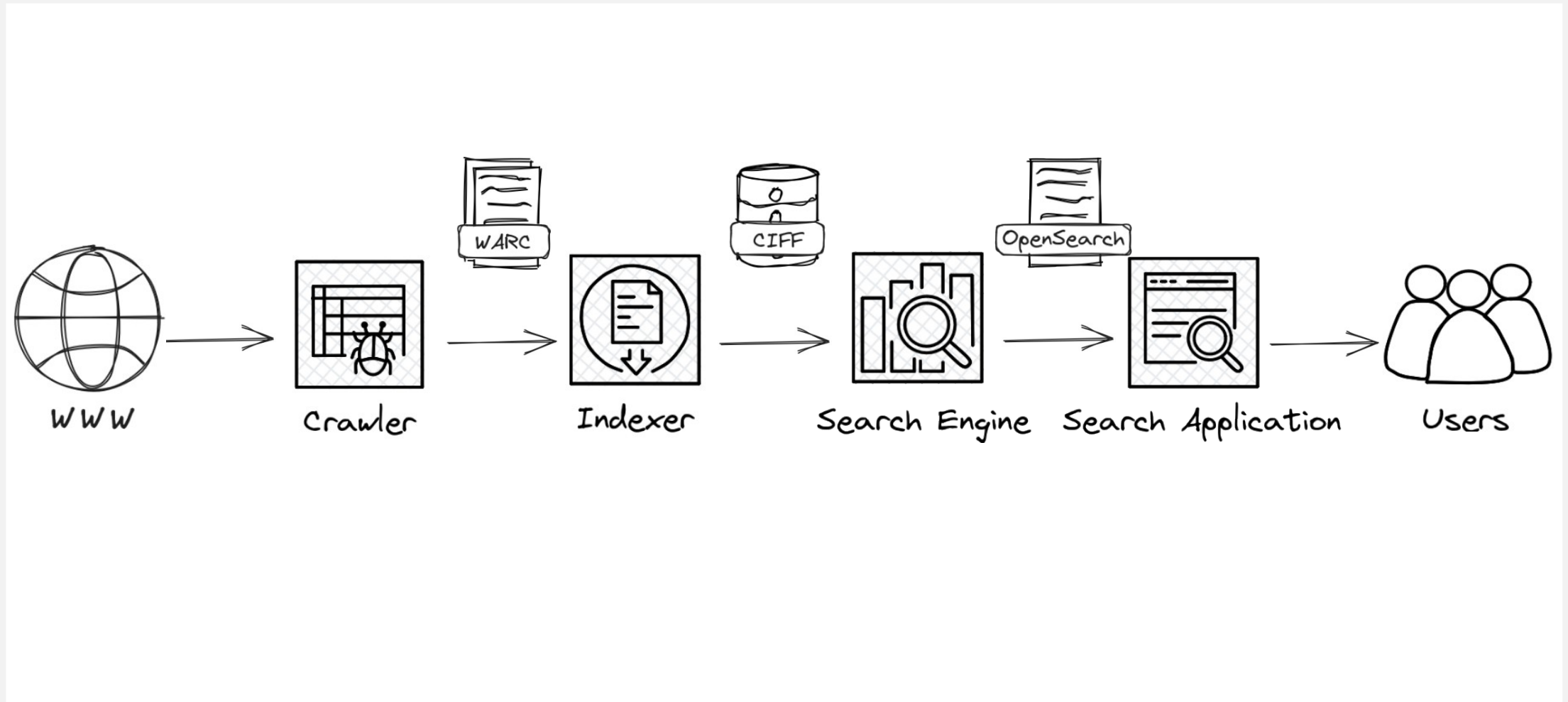
BUILDING SEARCH ENGINES COLLABORATIVELY



BUILDING SEARCH ENGINES COLLABORATIVELY



BUILDING SEARCH ENGINES COLLABORATIVELY



COMMON INDEX FILE FORMAT (CIFF)

- Inverted index
- For reproducibility of experiments
- Supported by:
 - Lucene (Anserini)
 - Terrier
 - PISA
 - JASSv2
 - OldDog
 - GeeseDB



CHALLENGE: TOKENIZATION

e.u. on-line 🔍

on-line
in the
E.U.

CHALLENGE: TOKENIZATION

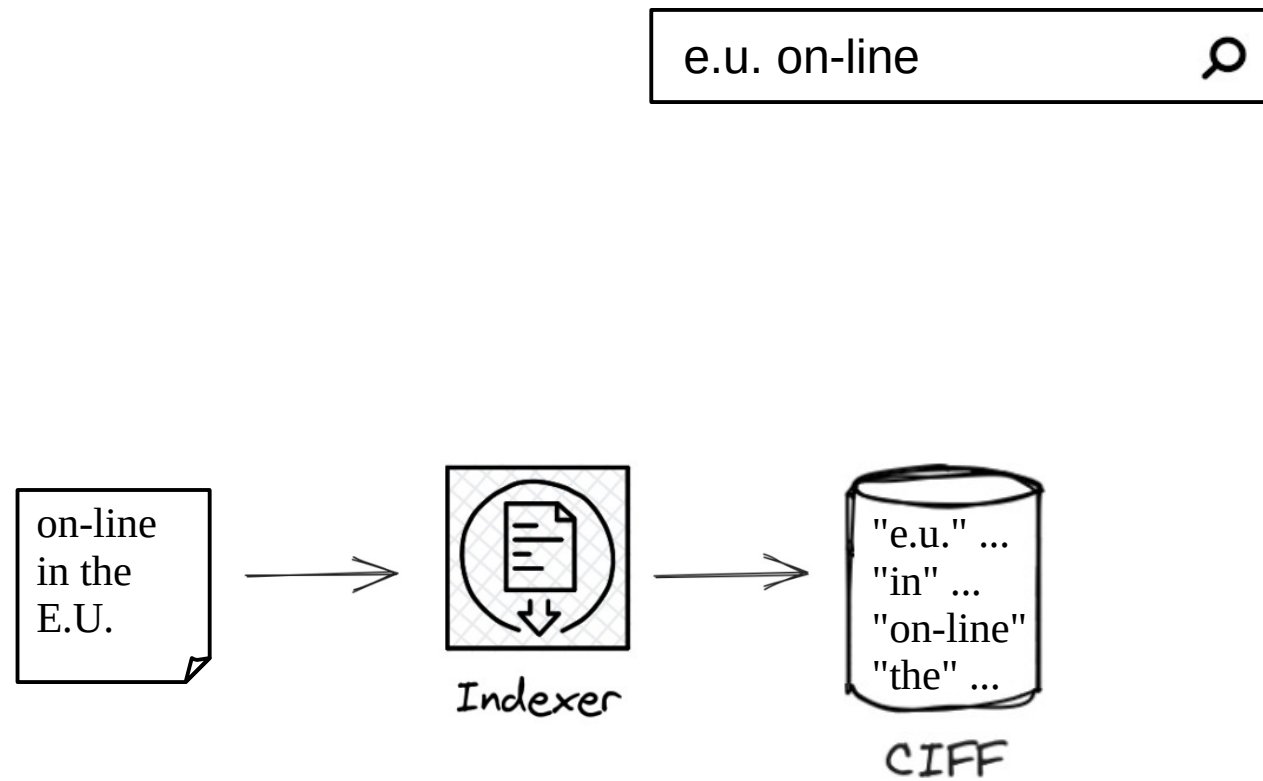
e.u. on-line 🔍

on-line
in the
E.U.

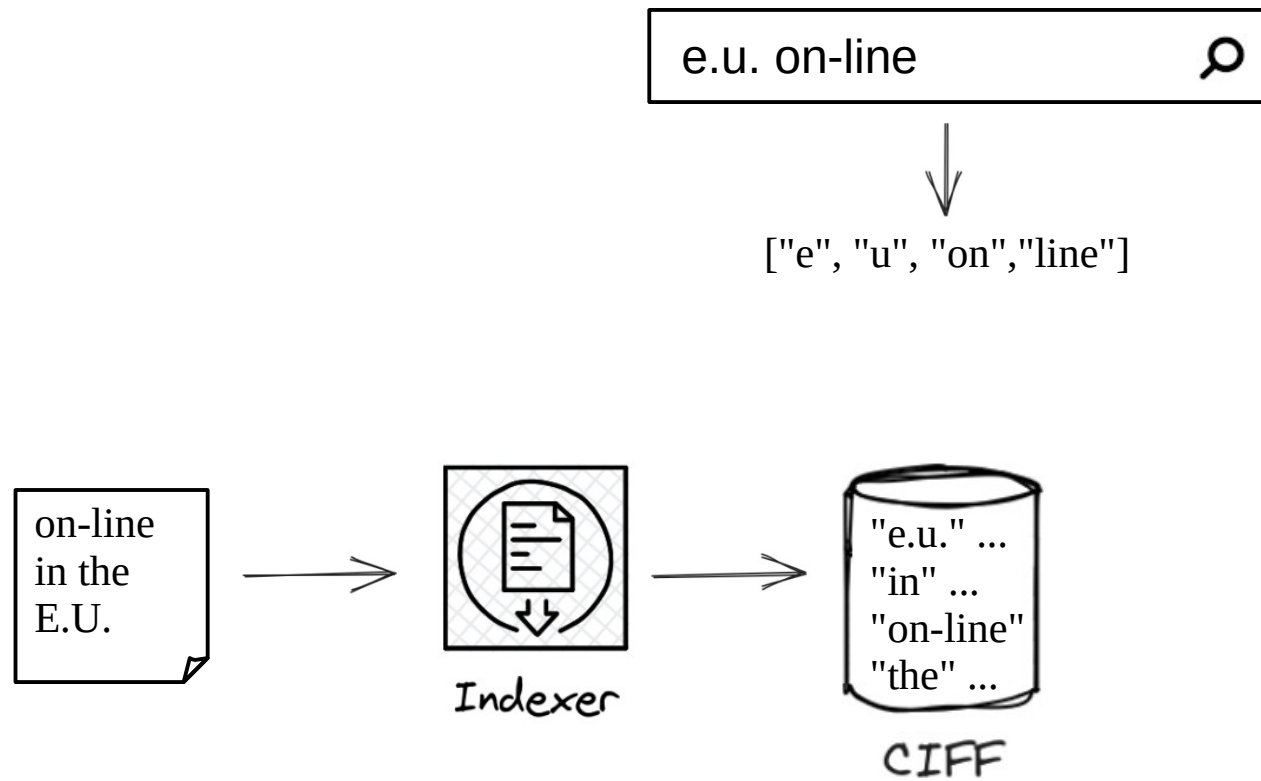


Indexer

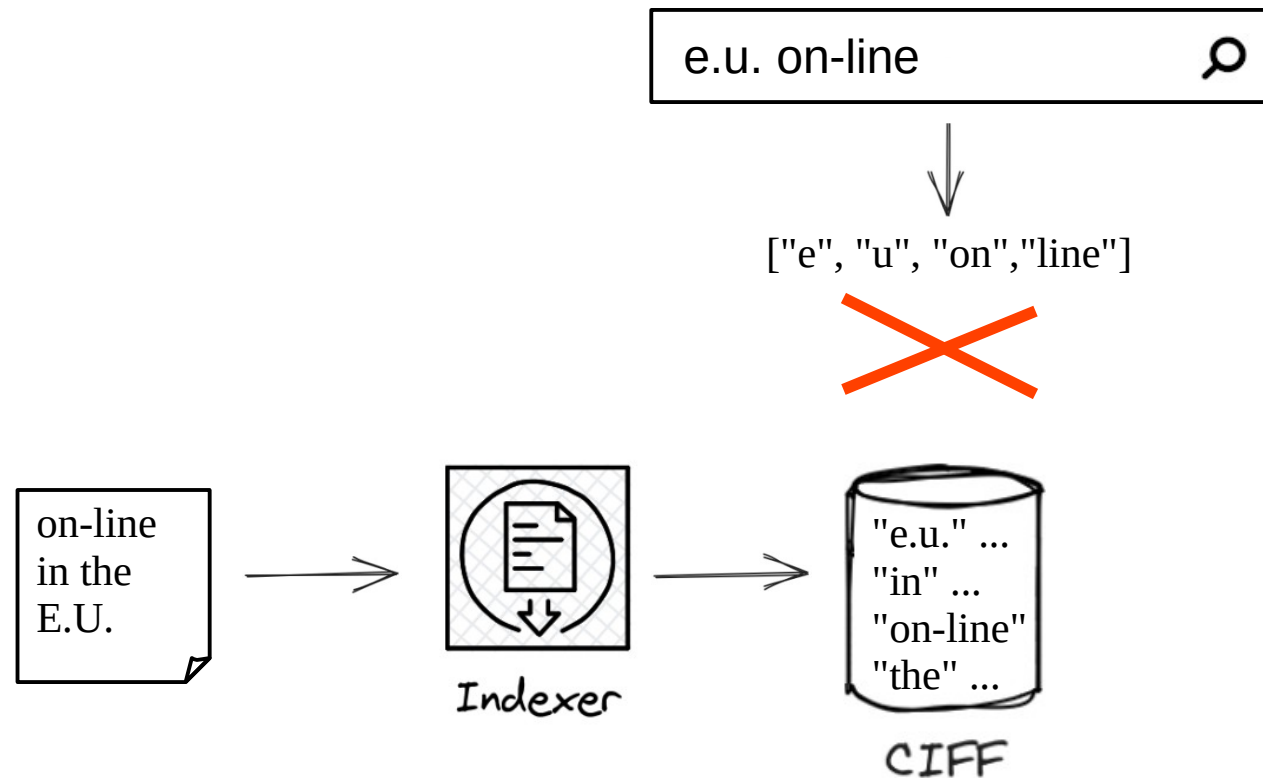
CHALLENGE: TOKENIZATION



CHALLENGE: TOKENIZATION



CHALLENGE: TOKENIZATION

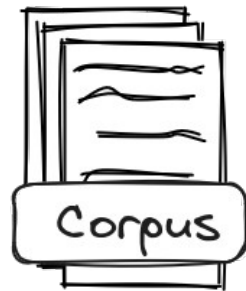


PROPOSED SOLUTION: A GENERIC TOKENIZER

- Drawing inspiration from LLM tokenizers

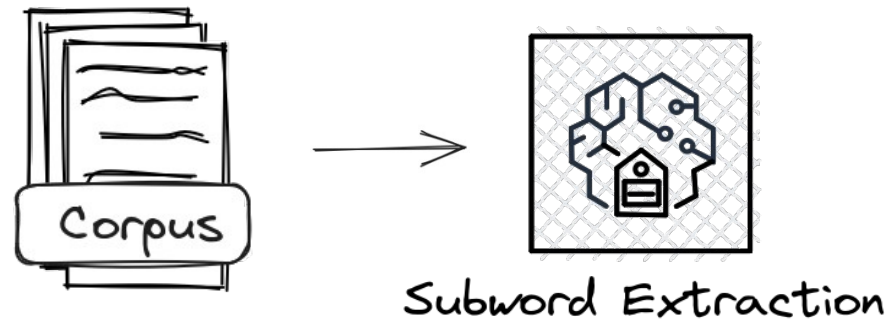
PROPOSED SOLUTION: A GENERIC TOKENIZER

- Drawing inspiration from LLM tokenizers



PROPOSED SOLUTION: A GENERIC TOKENIZER

- Drawing inspiration from LLM tokenizers



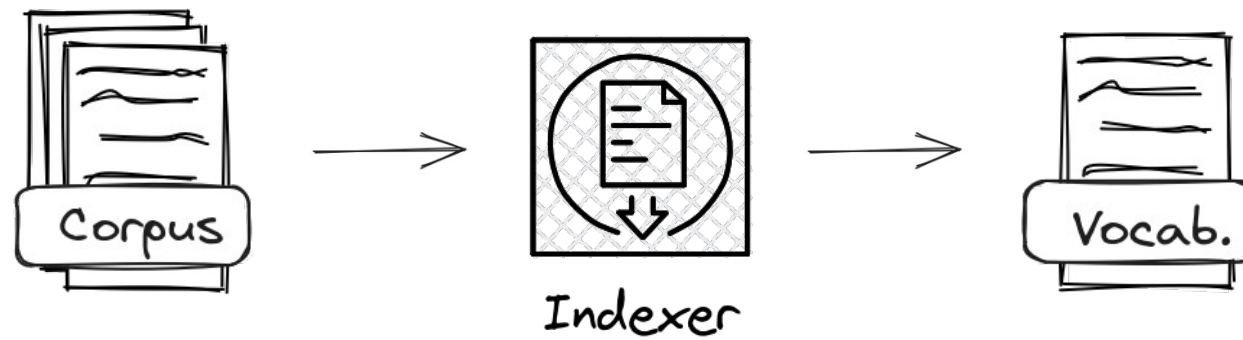
PROPOSED SOLUTION: A GENERIC TOKENIZER

- Drawing inspiration from LLM tokenizers

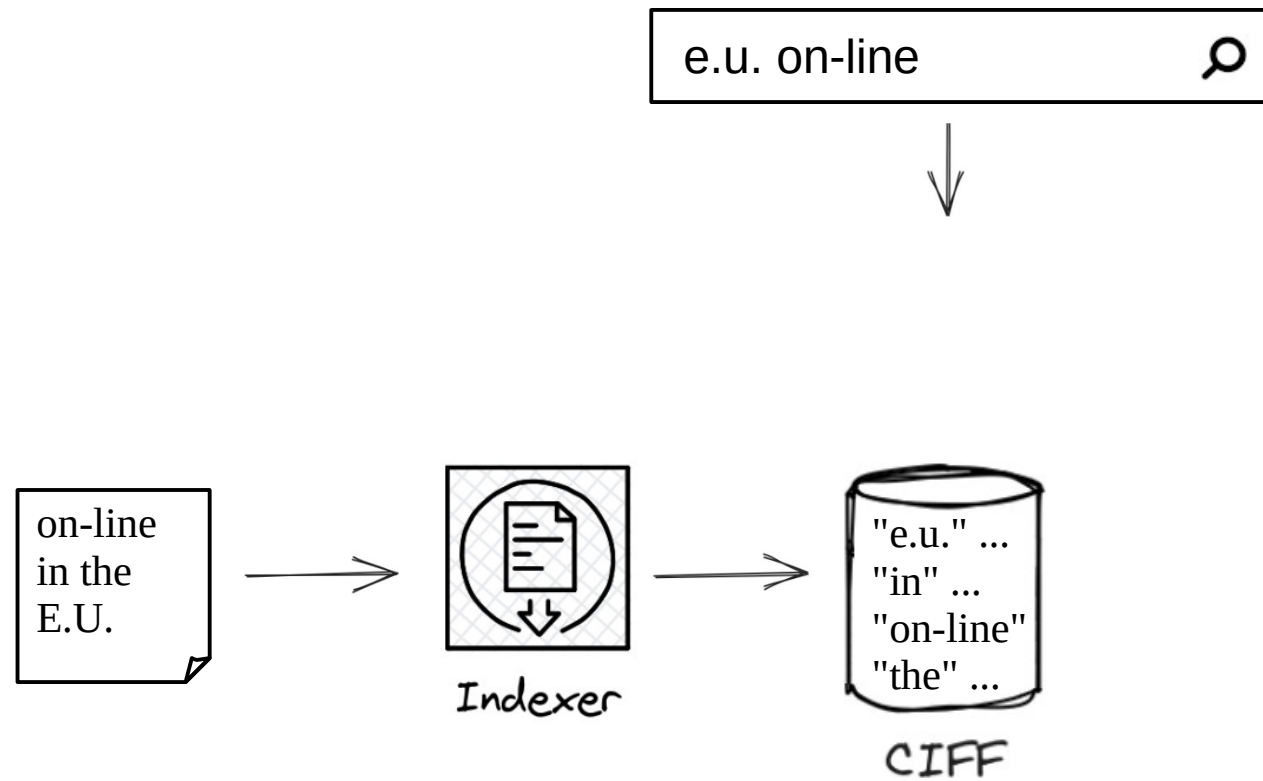


PROPOSED SOLUTION: A GENERIC TOKENIZER

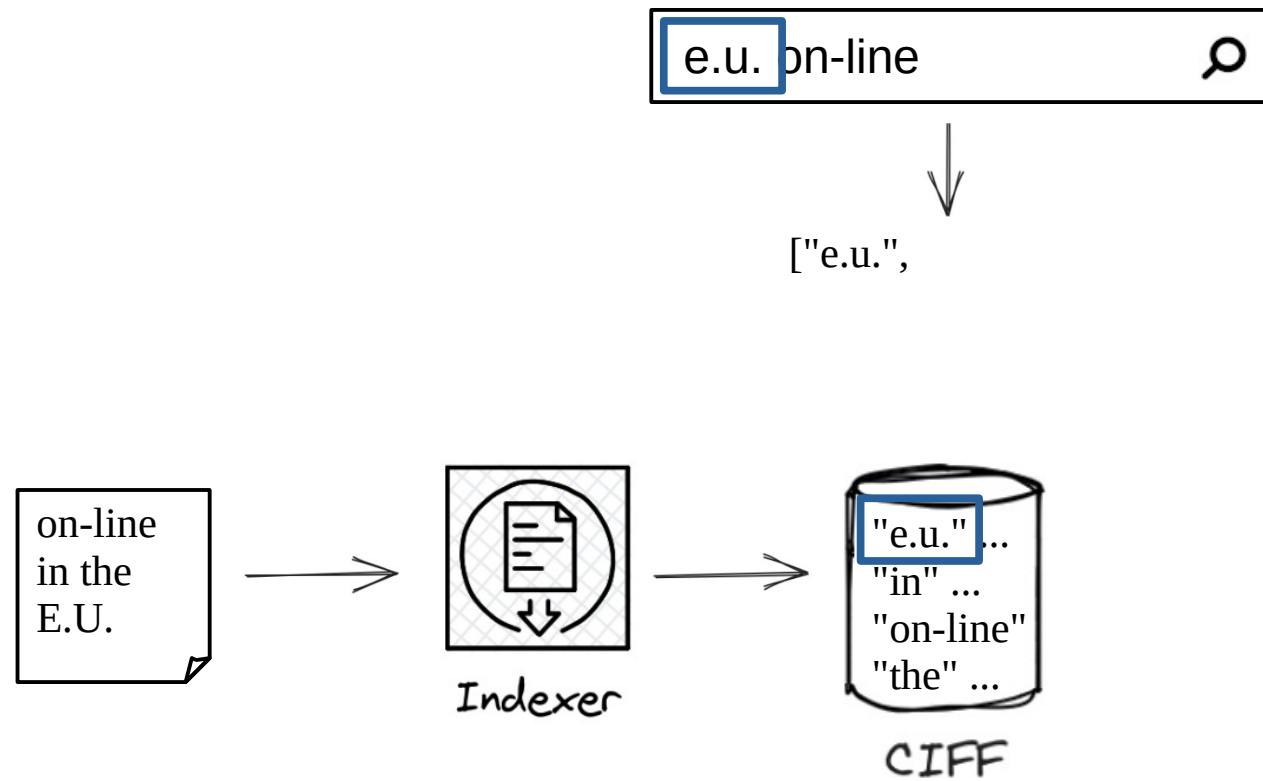
- Drawing inspiration from LLM tokenizers



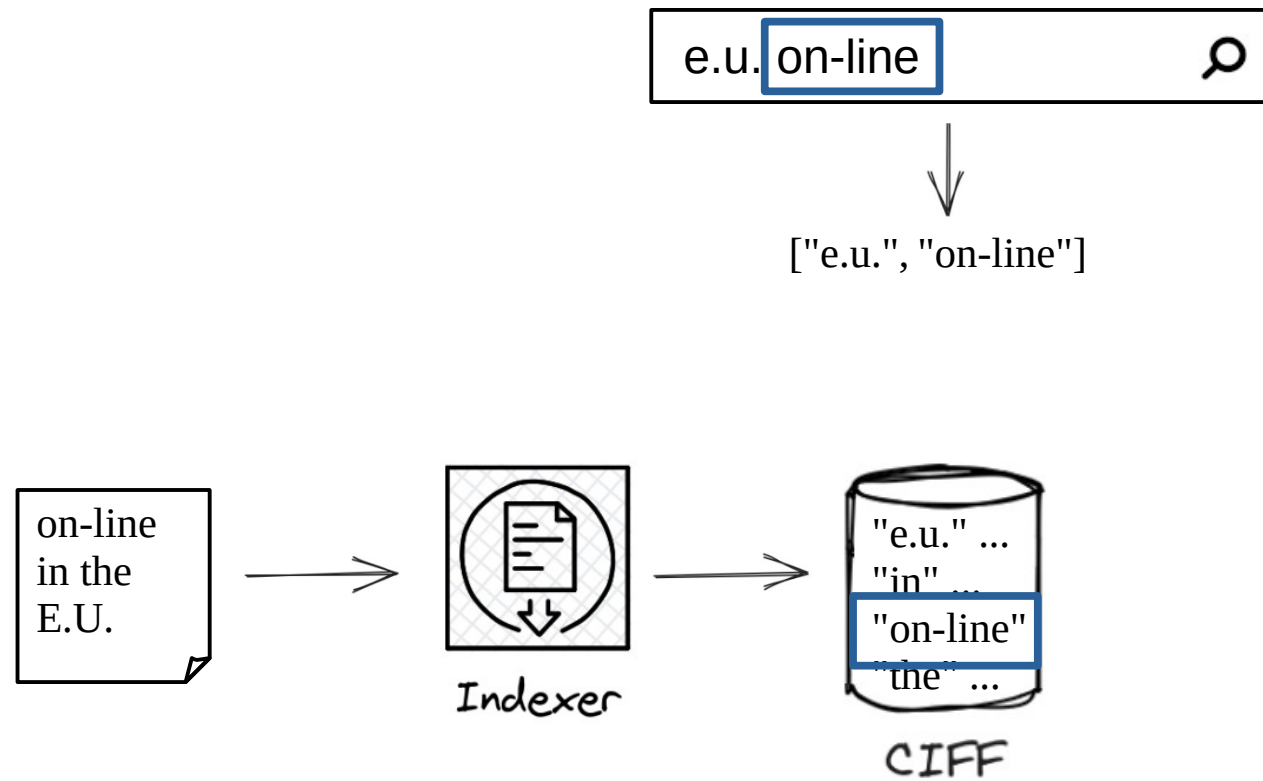
PROPOSED SOLUTION: A GENERIC TOKENIZER



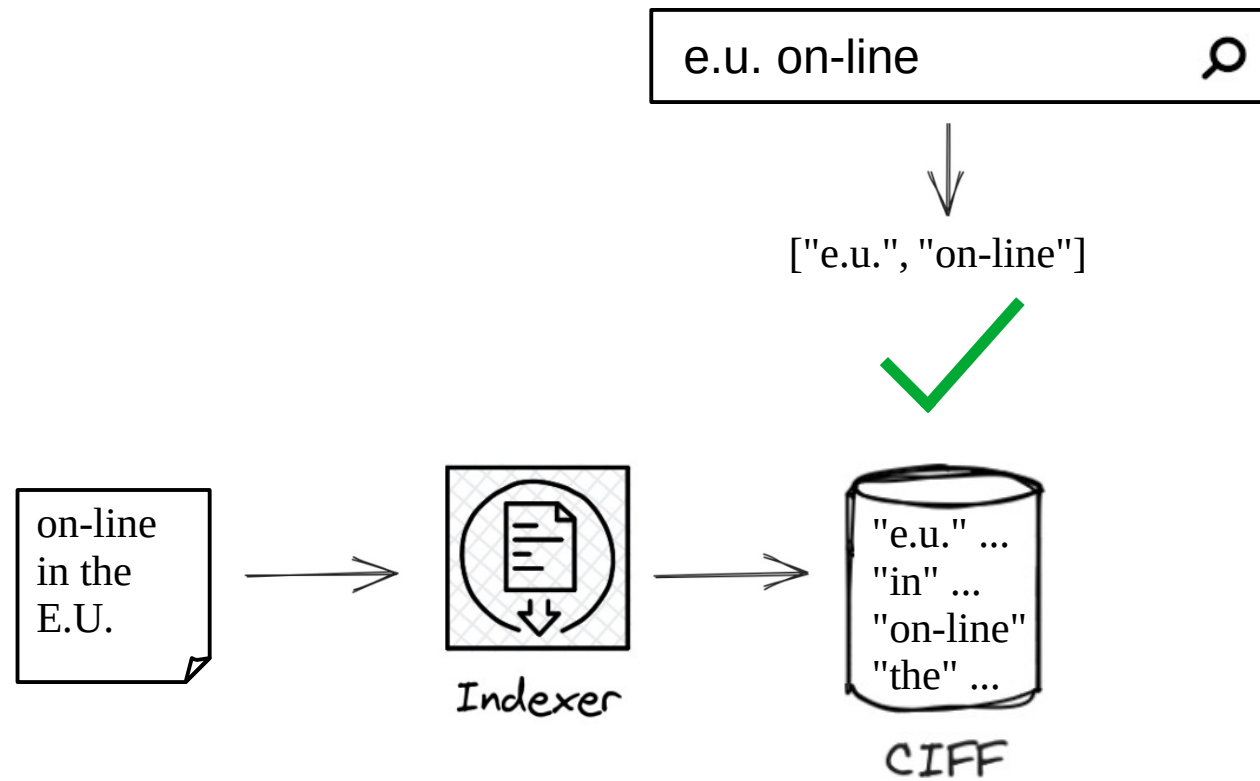
PROPOSED SOLUTION: A GENERIC TOKENIZER



PROPOSED SOLUTION: A GENERIC TOKENIZER



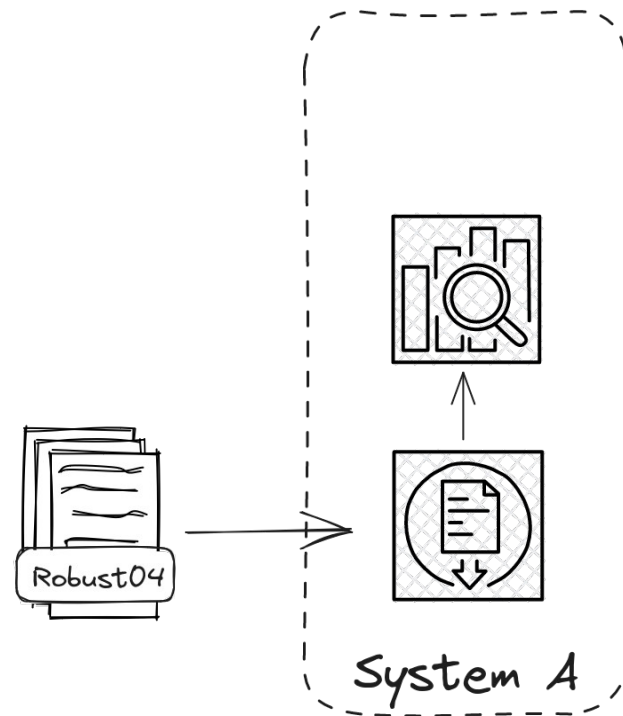
PROPOSED SOLUTION: A GENERIC TOKENIZER



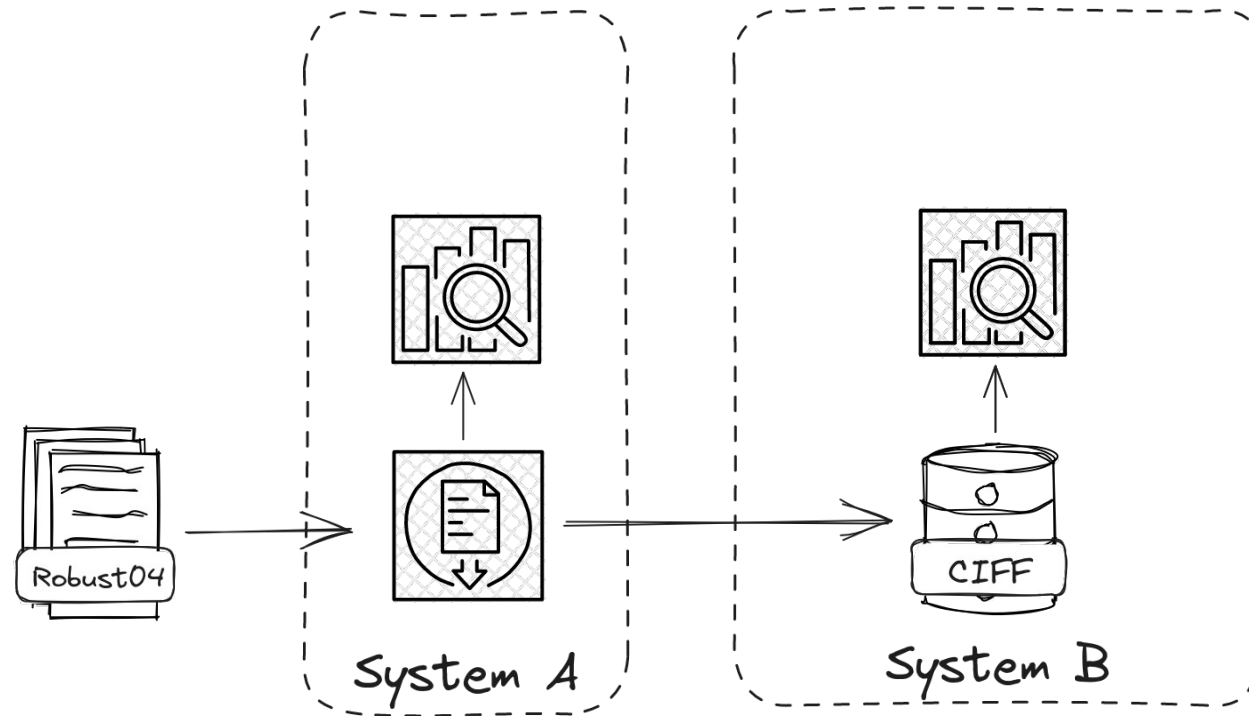
EVALUATING THE GENERIC CIFF TOKENIZER



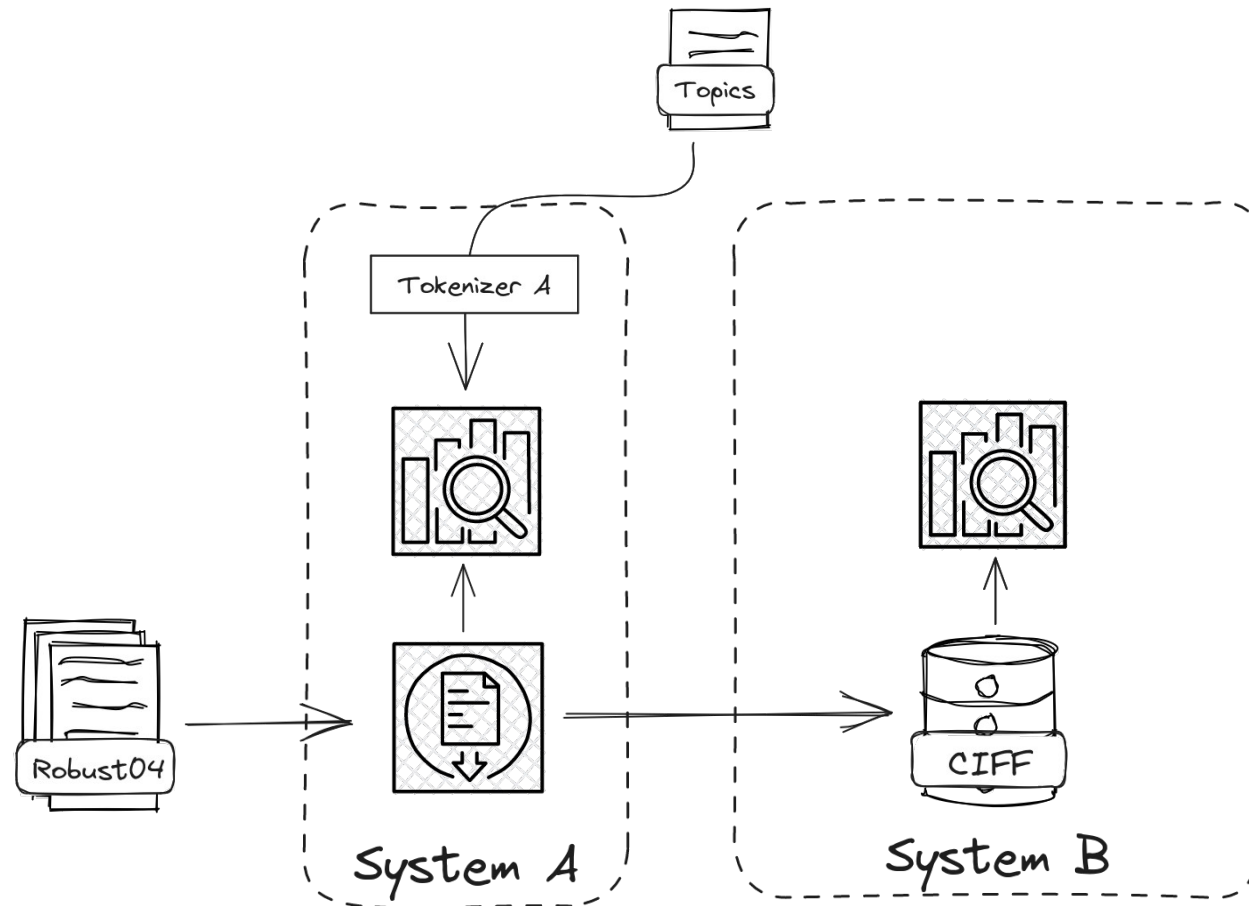
EVALUATING THE GENERIC CIFF TOKENIZER



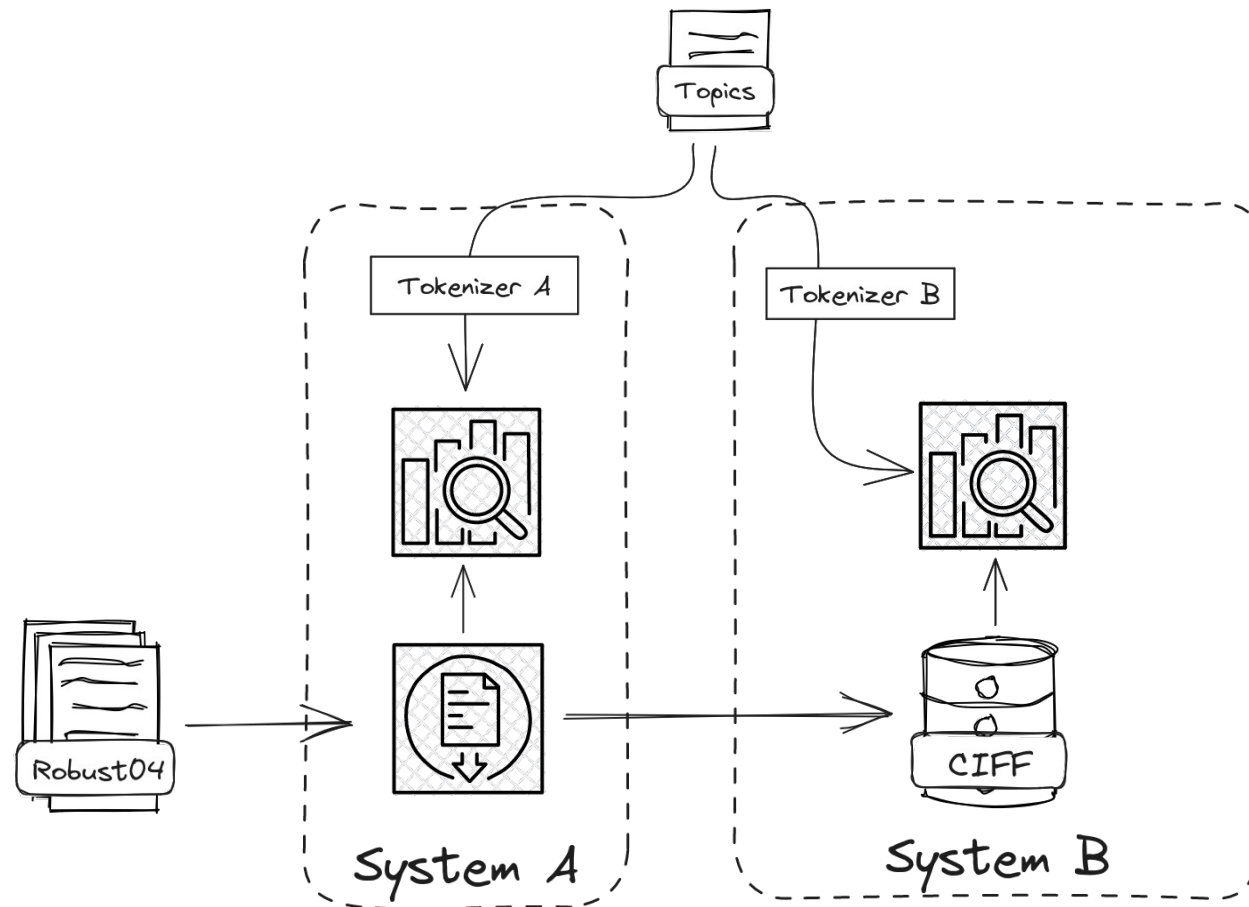
EVALUATING THE GENERIC CIFF TOKENIZER



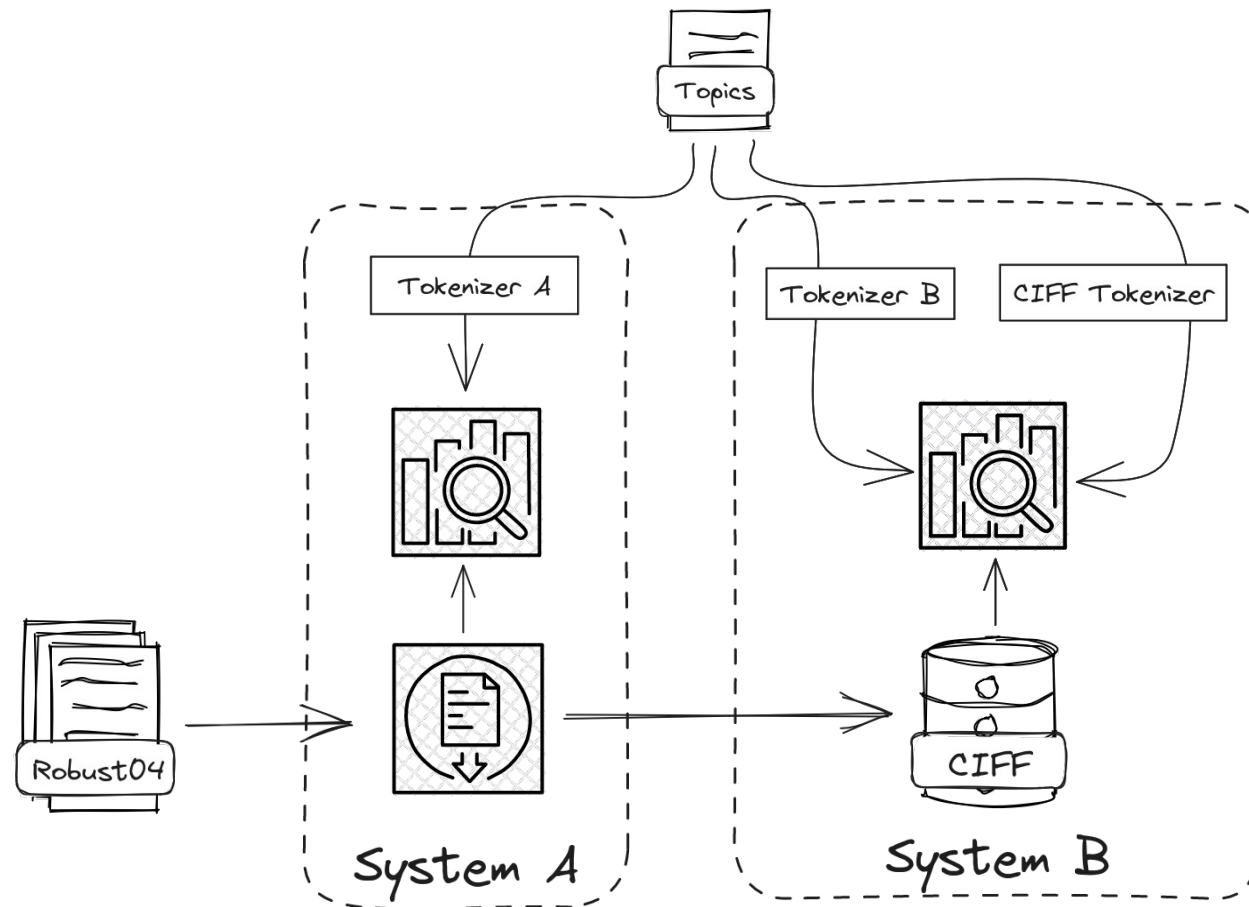
EVALUATING THE GENERIC CIFF TOKENIZER



EVALUATING THE GENERIC CIFF TOKENIZER



EVALUATING THE GENERIC CIFF TOKENIZER



RESULTS FOR TERRIER → GEESEDB

- Evaluated on all Robust04 topics

System	Tokenizer	MAP	nDCG
Terrier	Terrier	0.221 [†]	0.480 [†]
GeeseDB	NLTK	0.208	0.457
	CIFR	0.224[†]	0.482[†]

RESULTS FOR TERRIER → GEESEDB

- Evaluated on the Robust04 topics with hyphens or periods

System	Tokenizer	MAP	nDCG
Terrier	Terrier	0.234[†]	0.541[†]
GeeseDB	NLTK	0.081	0.292
	CIFFF	0.234[†]	0.541[†]

CONCLUSION

- Differences in tokenization cause drop in performance
 - Especially for punctuation-related tokenization
- A generic, dictionary-based tokenizer solves this

OTHER CHALLENGES

- Non-Western languages
- Stopwords
- Stemming
- Index updates